Biobank-wide association scan identifies risk factors for late-onset Alzheimer's disease and endophenotypes

Donghui Yan[1,#], Bowen Hu[2,#], Burcu F. Darst[3], Shubhabrata Mukherjee[4], Brian W. Kunkle[5], Yuetiva Deming[3], Logan Dumitrescu[6], Yunling Wang[1], Adam Naj[7], Amanda Kuzma[7], Yi Zhao[7], Hyunseung Kang[2], Sterling C. Johnson[8,9,10], Carlos Cruchaga[11], Timothy J. Hohman[6], Paul K. Crane[4], Corinne D. Engelman[3,8,10], Alzheimer's Disease Genetics Consortium (ADGC), Qiongshi Lu[2,12,*]

[1] University of Wisconsin-Madison, Madison, WI, 53706, USA

[2] Department of Statistics, University of Wisconsin-Madison, Madison, WI, 53706, USA

[3] Department of Population Health Sciences, University of Wisconsin-Madison, Madison, WI, 53726, USA

[4] Division of General Internal Medicine, Department of Medicine, University of Washington, Seattle, WA, 98195, USA

[5] University of Miami Miller School of Medicine, Miami, FL, 33136, USA

[6] Vanderbilt Memory and Alzheimer's Center, Vanderbilt University Medical Center, Vanderbilt University School of Medicine, Nashville, TN, 37212, USA

[7] School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA

[8] Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison, WI, 53719, USA

[9] Geriatric Research Education and Clinical Center, Wm. S. Middleton Memorial VA Hospital, Madison, WI, 53705, USA

[10] Alzheimer's Disease Research Center, University of Wisconsin School of Medicine and Public Health, Madison, WI, 53792, USA

[11] Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, 63110, USA

[12] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, 53792, USA

[#] These authors contributed equally to this work

* To whom correspondence should be addressed:

Dr. Qiongshi Lu

Department of Biostatistics and Medical Informatics

University of Wisconsin-Madison

425 Henry Mall, Room 2104

Madison, WI, USA 53706

qlu@biostat.wisc.edu

Keywords: GWAS, late-onset Alzheimer’s disease, risk factor, polygenic risk score

Abstract

Rich data from large biobanks, coupled with increasingly accessible association statistics from genome-wide association studies (GWAS), provide great opportunities to dissect the complex relationships among human traits and diseases. We introduce BADGERS, a powerful method to perform polygenic score-based biobank-wide association scans. Compared to traditional approaches, BADGERS uses GWAS summary statistics as input and does not require multiple traits to be measured in the same cohort. We applied BADGERS to two independent datasets for late-onset Alzheimer's disease (AD; N=61,212). Among 1,738 traits in the UK biobank, we identified 48 significant associations for AD. Family history, high cholesterol, and numerous traits related to intelligence and education showed strong and independent associations with AD. Further, we identified 41 significant associations for a variety of AD endophenotypes. While family history and high cholesterol were strongly associated with AD subgroups and pathologies, only intelligence and education-related traits predicted pre-clinical cognitive phenotypes. These results provide novel insights into the distinct biological processes underlying various risk factors for AD.

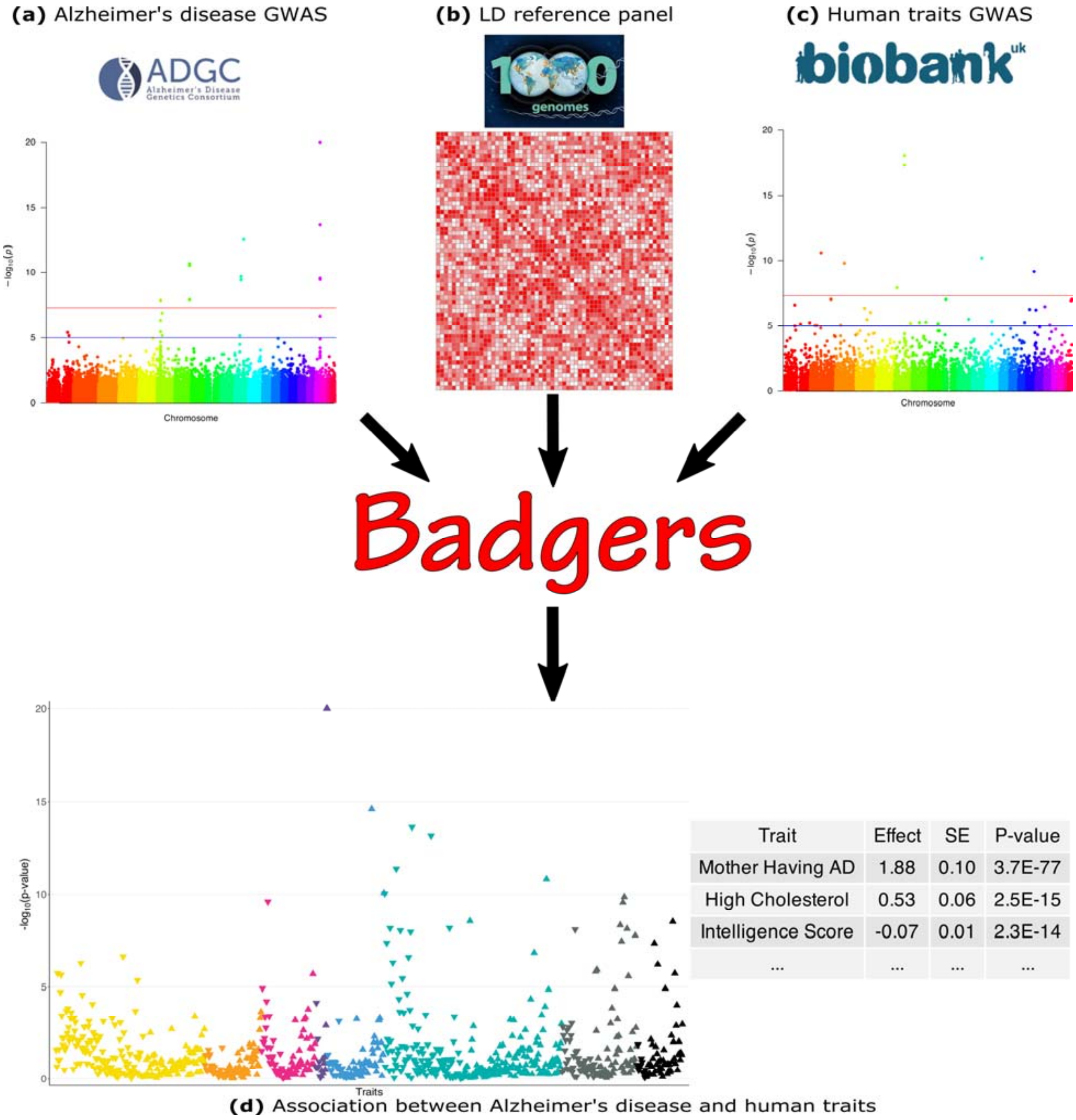
Introduction

Late-onset Alzheimer's disease (AD) is a prevalent, complex, and devastating neurodegenerative disease without a current cure. Millions of people are currently living with AD worldwide, and the number is expected to grow rapidly as the population continues to age [1, 2]. With the failure of numerous drug trials, it is of great interest to identify modifiable risk factors that can be potential targets in the therapeutics development for AD [3-5]. Epidemiological studies that directly test associations between measured risk factors and AD are difficult to conduct and interpret because identified associations are, in many cases, affected by confounding and reverse causality. Despite being ubiquitous challenges in risk factor studies for complex diseases, these issues are particularly critical for AD due to its extended pre-clinical stage – irreversible pathologic changes have already occurred in the decade or two prior to clinical symptoms [6]. On the other hand, Mendelian randomization methods [7-9] have been developed to identify causal risk factors for disease using data from genome-wide association studies (GWAS). Despite the favorable theoretical properties in identifying causal relationships, these methods have limited statistical power, thereby not suitable for hypothesis-free screening of risk factors.

Motivated by transcriptome-wide association study – an analysis strategy that identifies genes whose genetically regulated expression values are associated with disease [10-12], we seek a systematic and statistically powerful approach to identify risk factors using summary association statistics from large-scale GWAS. GWAS for late-onset AD has been successful, and dozens of associated loci have been identified to date [13-18]. Although direct information on risk factors is limited in these studies, dense genotype data on a large number of samples, in conjunction with independent reference datasets for thousands of complex human traits such as the UK biobank [19], make it possible to genetically impute potential risk factors and test their associations with AD. This strategy allows researchers to study risk factors that are not directly measured in AD studies. Furthermore, it reduces the reverse causality because the imputation models are trained on independent, younger, and mostly dementia-free reference cohorts, thereby improving the interpretability of findings.

Here, we introduce BADGERS (Biobank-wide Association Discovery using GEnetic Risk Scores), a statistically-powerful and computationally-efficient method to identify associations between a disease of interest and a large number of genetically-imputed complex traits using GWAS summary statistics. We applied BADGERS to identify associated risk factors for AD from 1,738 heritable traits in the UK biobank and replicated our findings in independent samples. Furthermore, we performed multivariate conditional analysis, Mendelian randomization, and follow-up association analysis with a variety of AD biomarkers, pathologies, and pre-clinical cognitive phenotypes to provide mechanistic insights into our findings.

Figure 1. BADGERS Workflow. BADGERS takes (a) Alzheimer's disease GWAS, (b) LD reference panel, and (c) Human traits GWAS from the UK biobank as input. The generated result will be the (d) Association between Alzheimer's disease and human traits. In graph (d), each triangle represents one human trait, and different colors represent different trait categories.



Trait	Effect	SE	P-value
Mother Having AD	1.88	0.10	3.7E-77
High Cholesterol	0.53	0.06	2.5E-15
Intelligence Score	-0.07	0.01	2.3E-14
...	...	...	...

Results

Method overview

Here, we briefly introduce the BADGERS model. The workflow of BADGERS is shown in **Figure 1**. A brief flowchart including all the analyses we contained in the manuscript was shown in the supplementary material (**Supplementary Figure 1)**. Complete statistical details are discussed in the **Methods** section. BADGERS is a two-stage method to test associations between traits. First, polygenic risk scores (PRS) are trained to impute complex traits using genetic data. Next, we test the association between a disease or trait of interest and various genetically-imputed traits. Given a PRS model, the imputed trait can be denoted as

$$\hat{T} = XW$$

where $X_{N \times M}$ is the genotype matrix for N individuals in a GWAS, and $W_{M \times 1}$ is the Mx1 matrix denotes pre-calculated weight values on SNPs in the PRS model. Then, we test the association between measured trait Y and imputed trait $\hat{T}$ via a univariate linear model

$$Y = \alpha + \hat{T}\gamma + \delta$$

The test statistic for γ can be expressed as:

$$Z = \frac{\hat{\gamma}}{se(\hat{\gamma})} \approx W^T \Gamma \tilde{Z}$$

where $\tilde{Z}$ is the vector of SNP-level association z-scores for trait Y, and Γ is a diagonal matrix with the jth diagonal element being the ratio between standard deviation of the jth SNP and that of $\hat{T}$.

This model can be further generalized to perform multivariate analysis. If K imputed traits are included in the analysis, we use a similar notation as in univariate analysis:

$$\hat{T}^* = XW^*$$

Here, $W^*_{M \times K}$ is a matrix and each column of W^* is the pre-calculated weight values on SNPs for each imputed trait. Then, the associations between Y and K imputed traits $\hat{T}_i$ $(1 \leq i \leq K)$ are tested via a multivariate linear model

$$Y = \alpha^* + \hat{T}^*\gamma^* + \delta^*$$

where $\gamma^* = (\gamma_1, \dots, \gamma_K)^T$ is the vector of regression coefficients. The z-score for γ_i $(1 \leq i \leq K)$ can be denoted as:

$$Z_i = \frac{\hat{\gamma}_i}{se(\hat{\gamma}_i)} \approx \frac{1}{\sqrt{U_{ii}}} I_i^T U (W^*)^T \Theta \tilde{Z}$$

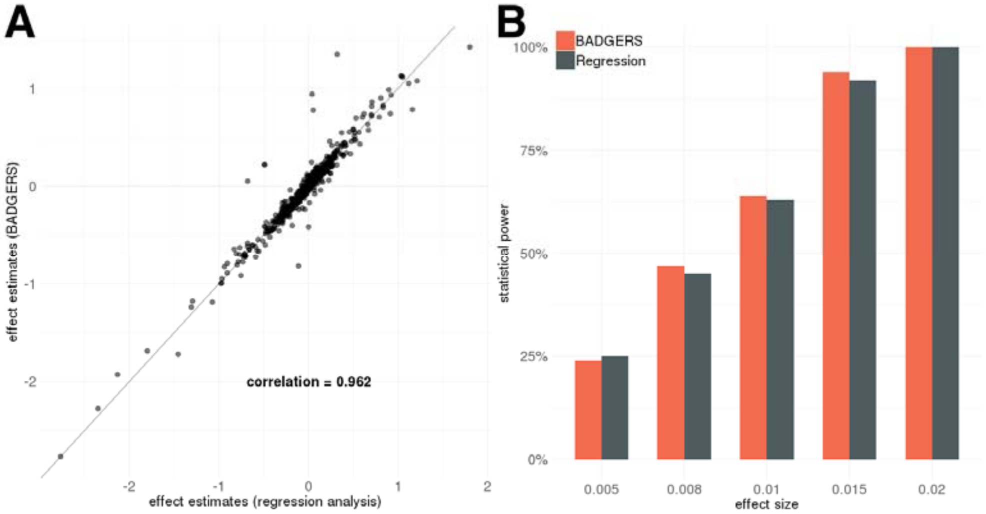
where U is the inverse variance-covariance matrix of $\hat{T}^*$; I_i is the $K \times 1$ vector with the ith element being 1 and all other elements equal to 0; Θ is a $M \times M$ diagonal matrix with the ith diagonal element being $\sqrt{var(X_i)}$; and $\tilde{Z}$ is defined the same as the univariate case as the vector of SNP-level association z-scores for trait Y.

Simulations

We used real genotype data from the Genetic Epidemiology Research on Adult Health and Aging (GERA) to conduct simulation analyses (**Methods**). First, we evaluated the performance of our method on data simulated under the null hypothesis. We tested the associations between randomly simulated traits and 1,738 PRS from the UK biobank and did not observe inflation in type-I error (**Supplementary Table 1**). Similar results were also observed when we simulated traits that are heritable but not directly associated with any PRS. Since BADGERS only uses summary association statistics and externally estimated linkage disequilibrium (LD) as input, we also compared effect estimates in BADGERS with those of traditional regression analysis based on individual-level data. Regression coefficient estimates and association p-values from these two methods were highly consistent in both simulation settings (**Figure 2A** and **Supplementary Figures 2-4)**, showing minimal information loss in summary statistics compared to individual-level data indicating highly consistent performance compared to methods based on individual-level data. To evaluate the statistical power of BADGERS, we simulated traits by combining effects from randomly selected PRS and a polygenic background (**Methods**). We set the effect size of PRS to be 0.02, 0.015, 0.01, 0.008, and 0.005. BADGERS showed comparable statistical power to the regression analysis based on individual-level genotype and phenotype data (**Figure 2B, Supplementary Table 2)**. Overall, our results suggest that using summary association statistics and externally estimated LD as a proxy for individual-level genotype and phenotype data does not inflate type-I error rate or decrease power. The performance of BADGERS is comparable to regression analysis based on individual-level data. We also studied if more sophisticated polygenic risk prediction methods could potentially lead to higher statistical power in downstream association tests. We compared the performance of PRS based on marginal effect sizes with that of LDpred, a more sophisticated PRS model that jointly estimates SNP effects via a Bayesian framework [20]. Imputation models based on multivariate analysis indeed improved the results. When using marginal PRS to impute traits, the correlation between

$\hat{\gamma}_i$ and γ_i was 0.79. This correlation improved to 0.91 when using LDpred PRS **(Supplementary Figure 5)**. However, such improvement did not substantially affect the statistical power in association testing. Using marginal PRS, our analysis achieved a statistical power of 86% to identify associations at a type-I error rate of 0.05, and the power was 88% when using multivariate effect estimates to calculate PRS. These results suggest that while more sophisticated PRS methods can improve the results in BADGERS, simple PRS based on marginal effects also shows reasonably good performance.

Figure 2. Simulation results. BADGERS and regression analysis based on individual-level data showed **(A)** highly consistent effect size estimates for 1,738 PRS in simulation and **(B)** comparable statistical power (setting 3).

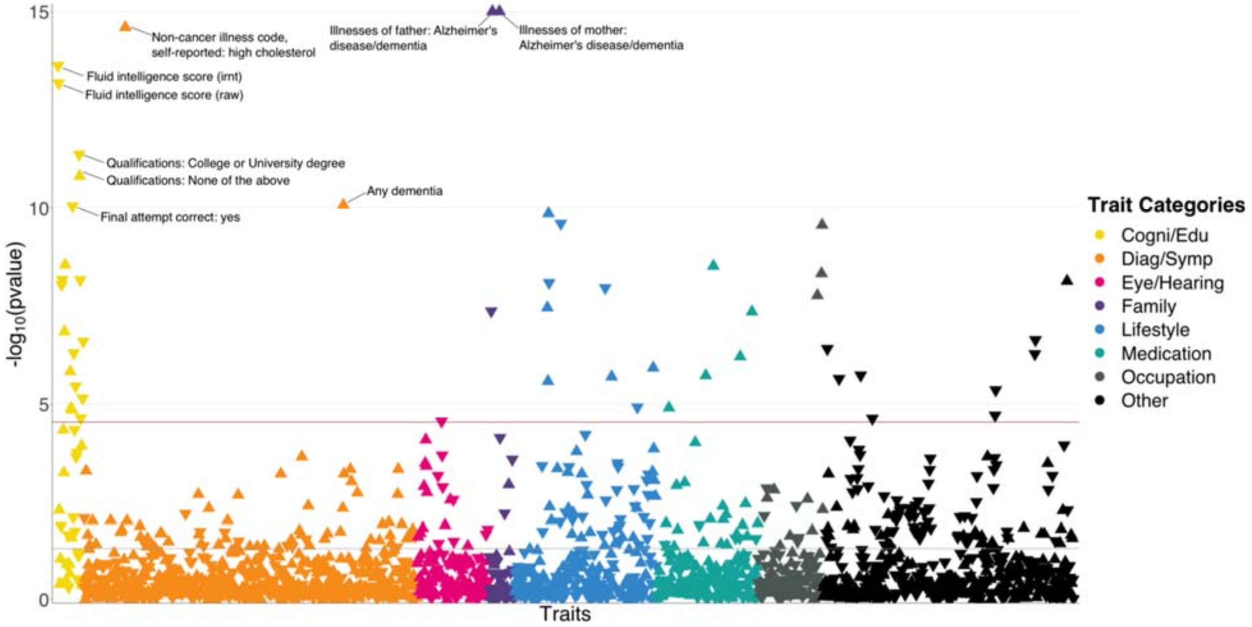


Identify risk factors for late-onset AD among 1,738 heritable traits in the UK biobank

We applied BADGERS to conduct a biobank-wide association scan (BWAS) for AD risk factors from 1,738 heritable traits (p<0.05; **Methods**) in the UK biobank. We repeated the analysis on two independent GWAS datasets for AD and further combined the statistical evidence via meta-analysis (**Supplementary Figure 6**). We used stage-I association statistics from the International Genomics of Alzheimer's Project (IGAP; N=54,162) as the discovery phase, then replicated the findings using 7,050 independent samples from the Alzheimer's Disease Genetics Consortium (ADGC). We identified 50 significant trait-AD associations in the discovery phase after correcting for multiple testing, among which 14 had p<0.05 in the replication analysis. Despite the considerably smaller sample size in the replication phase, top traits identified in the discovery stage showed strong enrichment for p<0.05 in the replication analysis (enrichment=2.5, p=2.2e-5; hypergeometric test). In the meta-analysis, a total of 48 traits reached Bonferroni-corrected statistical significance and showed consistent effect directions in the discovery and replication analyses (**Figure 3** and **Supplementary Table 3**).

Unsurprisingly, many identified associations were related to dementia and cognition. Family history of AD and dementia showed the most significant associations with AD (p=3.7e-77 and 5.2e-28 for illnesses of mother and father, respectively). Having any dementia diagnosis is also strongly and positively associated (p=8.5e-11). In addition, we observed consistent and negative associations between better performance in cognition test and AD risk. These traits include fluid intelligence score (p=2.4e-14), time to complete round in cognition test (p=2.8e-9), correct final attempt (p=9.1e-11), and many others. Consistently, education attainment showed strong associations with AD. Age completed full time education (p=2.5e-7) was associated with lower AD risk. Four out of seven traits based on a survey about education and qualifications were significantly associated with AD (**Supplementary Figure 7**). Higher education such as having a university degree (p=4.4e-12), A levels/AS levels or equivalent (p=6.9e-9), and professional qualifications (p=7.1e-6) were associated with lower AD risk. In contrast, choosing "none of the above" in this survey was associated with a higher risk (p=1.6e-11). Other notable strong associations include high cholesterol (p=2.5e-15; positive), lifestyle traits such as cheese intake (p=2.5e-10; negative), occupation traits such as job involving heavy physical work (p=2.7e-10; positive), anthropometric traits including height (p=5.3e-7; negative), and traits related to pulmonary function, e.g. forced expiratory volume in 1 second (FEV1; p=1.9e-6; negative). Detailed information on all associations is summarized in **Supplementary Table 3.**

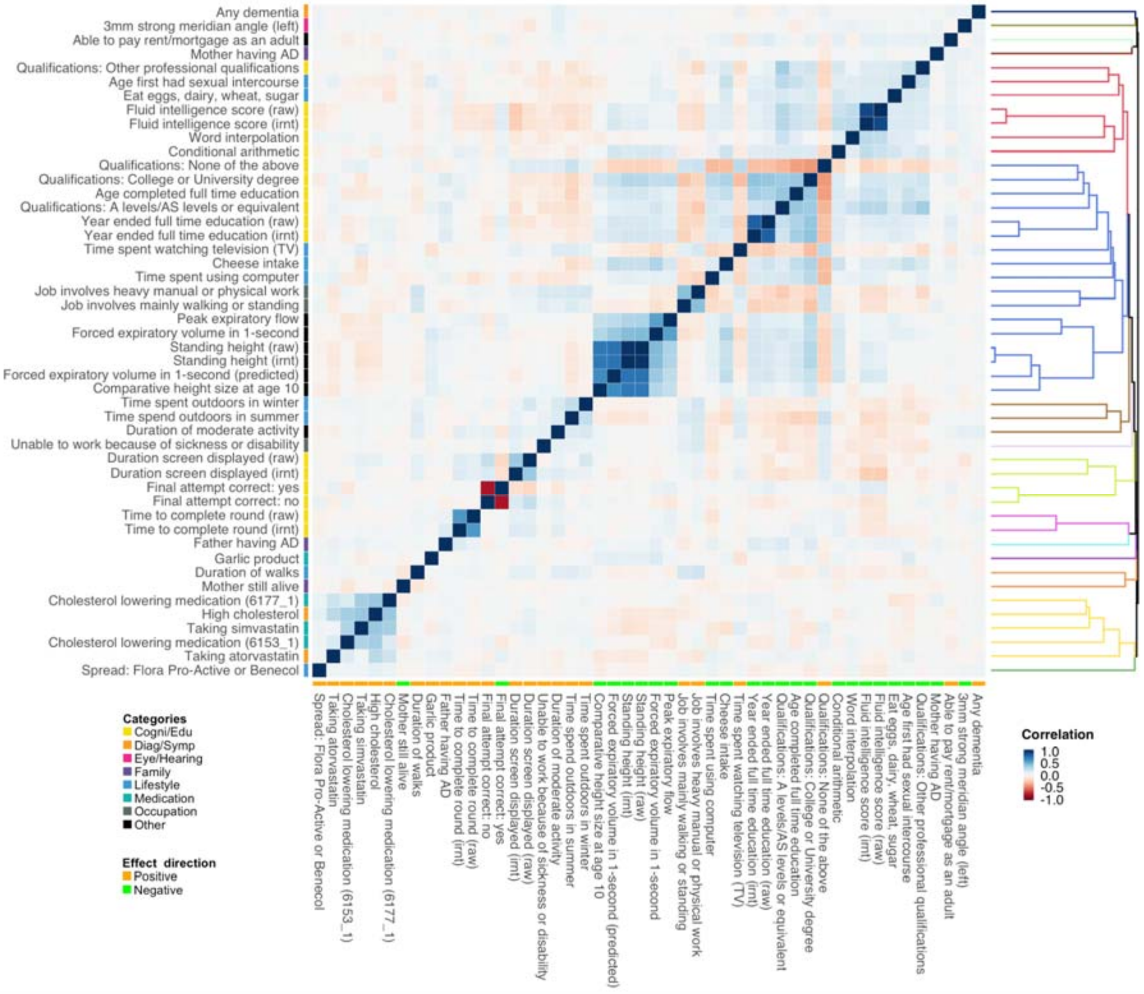
Figure 3. PRS-based BWAS identifies risk factors for AD. Meta-analysis p-values for 1,738 heritable traits in the UK biobank are shown in the figure. P-values are truncated at 1e-15 for visualization purposes. The horizontal line marks the Bonferroni-corrected significance threshold (i.e. p=0.05/1738). Positive associations point upward, and negative associations point downward.



Multivariate conditional analysis identifies independently associated risk factors

Of note, associations identified in the marginal analysis are not guaranteed to be independent. We observed clear correlational structures among the identified traits (**Figure 4**). For example, PRS of various intelligence and cognition-related traits are strongly correlated, and consumption of cholesterol-lowering medication is correlated with self-reported high cholesterol. To account for the correlations among traits and identify risk factors that are independently associated with AD, we performed multivariate conditional analysis using GWAS summary statistics (**Methods**). First, we applied hierarchical clustering to the 48 traits we identified in marginal association analysis and divided these traits into 15 representative clusters. The traits showing the most significant marginal association in each cluster were included in the multivariate analysis (**Supplementary Figure 8**). Similar to the marginal analysis, we analyzed IGAP and ADGC data separately and combined the results using meta-analysis (**Supplementary Table 4**). All 15 representative traits remained nominally significant (p<0.05) and showed consistent effect directions between marginal and conditional analyses (**Supplementary Table 5**). However, several traits showed substantially reduced effect estimates and inflated p-values in multivariate analysis, including fluid intelligence score, mother still alive, unable to work because of sickness or disability, duration of moderate activity, and intake of cholesterol-lowering spread. Interestingly, major trait categories that showed the strongest marginal associations with AD (i.e. family history, high cholesterol, and education/cognition) were independent from each other. Paternal and maternal family history also showed independent associations with AD, consistent with the low correlation between their genetic risk scores (correlation= 0.052).

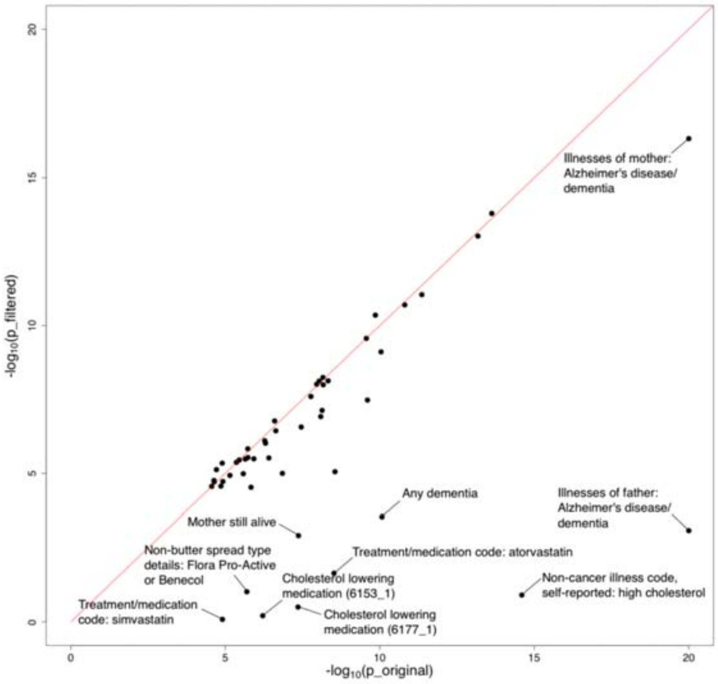
Figure 4. PRS correlation matrix for the 48 traits identified in marginal association analysis. Trait categories and association directions with AD are annotated. The dendrogram indicates the results of hierarchical clustering. We used 1000 Genomes samples with European ancestry to calculate PRS and evaluate their correlations. Label “irnt” means that trait values were standardized using rank-based inverse normal transformation in the GWAS analysis.



Influence of the *APOE* region on identified associations

Further, we evaluated the impact of *APOE* on identified associations. We removed the extended *APOE* region (chr19: 45,147,340–45,594,595; hg19) from summary statistics of the 48 traits showing significant marginal associations with AD and repeated the analysis. We observed a substantial drop in the significance level of many traits, especially family history of AD, dementia diagnosis, and high cholesterol (**Figure 5, Supplementary Figure 9,** and **Supplementary Table 6**). 38 out of 48 traits remained significant under stringent Bonferroni correction after *APOE* removal. Interestingly, the associations between AD and almost all cognition/intelligence traits were virtually unchanged, suggesting a limited role of *APOE* in these associations.

Figure 5. Influence of the *APOE* region on trait-AD associations. The horizontal and vertical axes denote association p-values before and after removal of the *APOE* region, respectively. Original p-values (i.e. the x-axis) were truncated at 1e-20 for visualization purposes.



Causal inference via Mendelian randomization

Next, we investigated the evidence for causality among identified associations. We performed Mendelian randomization (MR-IVW; **Methods**) in IGAP and ADGC datasets separately and meta-analyzed the results on the complete set of 1,738 heritable traits from the UK biobank. A total of 48 traits reached Bonferroni-corrected statistical significance and showed consistent effect directions in the discovery and replication analyses using BADGERS. In contrast, MR-IVW only identified 9 traits with Bonferroni-corrected statistical significance. Among these 9 traits, 7 were also identified by BADGERS (**Supplementary Table 7**). The signs of all significant causal effects identified by MR-IVW were consistent with results from BADGERS. The most significant effect was family history (p=1.1e-233 and 1.7e-69 for maternal and paternal history, respectively). Dementia diagnosis (p=9.1e-7), high cholesterol (p=4.1e-6), A levels/AS levels education (p=1.7e-4), and time spent watching television (p=2.4e-4) were also among the top significant effects. Of note, the fluid intelligence score, one of the most significant associations identified by BADGERS, did not reach statistical significance in MR (p=0.06), which may be explained by its polygenic genetic architecture. It is also worth noting that if we scan all 1,738 traits using BADGERS and then apply MR-IVW on the 48 Bonferroni-corrected significant traits, 23 could reach nominal significance (p<0.05) in MR, and 7 could reach significance under Bonferroni correction (p<0.05/48; **Supplementary Table 8**).

We also compared BADGERS with another more recent method GSMR [9]. Due to the smaller sample size in the ADGC dataset, we only applied GSMR to the IGAP summary statistics. In total, 18 traits reached statistical significance under Bonferroni correction (**Supplementary Table 9**). However, these results showed only moderate consistency with MR-IVW and BADGERS. Among the 18 significant traits, only 1 trait, maternal family history of Alzheimer's disease and dementia, overlapped with significant traits identified by MR-IVW. Six out of 18 traits overlapped with significant traits identified by BADGERS. Among the 18 significant traits, 8 are related to body fat mass and 2 are related to educational attainment. The most significant effect was illnesses of mother (p=2.4e-294). College or University degree (p=4.84e-6), education; none of the above(p=3.6e-4), A levels/AS levels education (p=3.8e-6), and time spent watching television (p=4.0e-3) were also among top significant effects. Notably, GSMR failed to identify paternal family history or high cholesterol as risk factors for Alzheimer's disease. If we only consider the 48 significant traits identified by BADGERS, 11 were nominally significant (p<0.05). However, 23 traits did not have enough significant SNPs to perform the

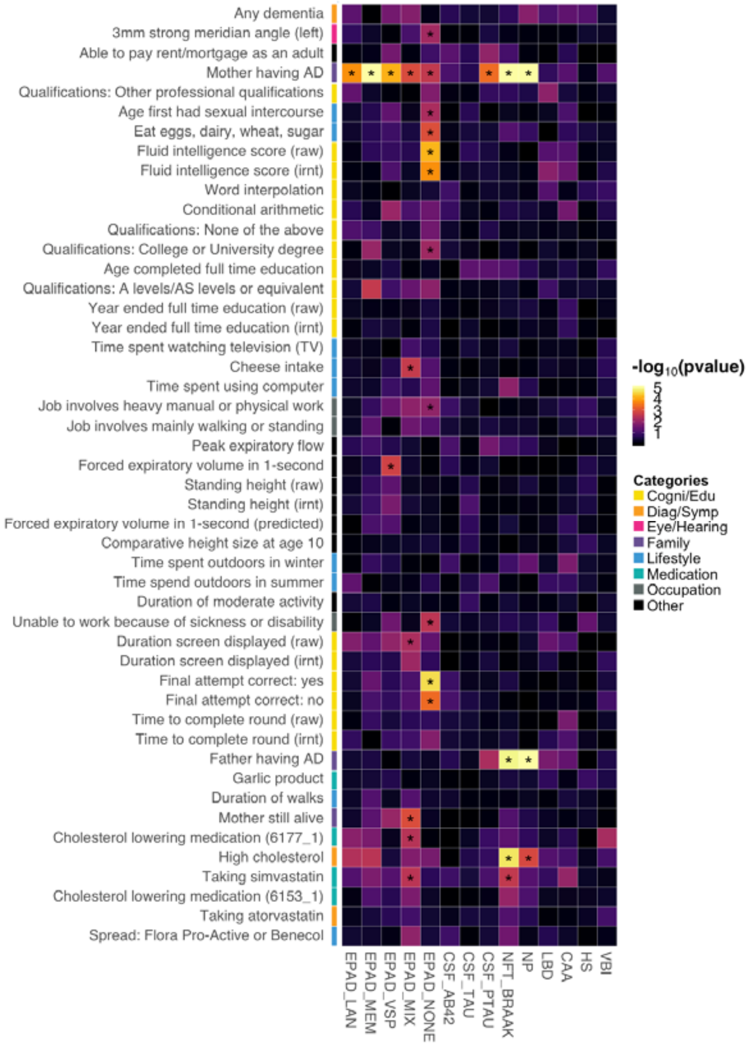
GSMR analysis (at least 10 SNPs are required). The signs of all significant causal effects identified by GSMR were consistent between association effects in BADGERS.

Additionally, we included GSMR analysis results after removing *APOE* region from the 48 identified traits. Only maternal family history reached Bonferroni-corrected statistical significance, further demonstrating the lack of statistical power in MR when performing biobank-wide scans (**Supplementary Table 10**).

Associations with AD subgroups, biomarkers, and pathologies

To further investigate the mechanistic pathways for the identified risk factors, we applied BADGERS to a variety of AD subgroups, biomarkers, and neuropathologic features (**Supplementary Table 11**). Overall, 29 significant associations were identified under a false discovery rate (FDR) cutoff of 0.05, and these endophenotypes showed distinct association patterns with AD risk factors (**Figure 6**; **Supplementary Figure 10**). First, we tested the associations between the 48 AD-associated traits and five AD subgroups defined in the Executive Prominent Alzheimer's Disease (EPAD) study, i.e. memory, language, visuospatial, none, and mix (**Methods**) [21, 22]. Maternal family history of AD and dementia was strongly and consistently associated with all five EPAD subgroups (**Supplementary Table 12**), with memory subgroup showing the strongest association (p=3.3e-16), which is consistent with the higher frequency of *APOE* ε4 in this subgroup [21]. Paternal family history was not strongly associated with any subgroups, but the effect directions were consistent. Interestingly, intelligence and cognition-related traits such as correct final attempt in cognitive test (p=2.7e-5) and fluid intelligence score (p=6.8e-5) were specifically associated with the "none" subgroup – AD samples without relative impairment in any of the four cognitive domains. High cholesterol and related traits were associated with language, memory, and mix (i.e. AD samples with relative impairment in two or more domains) subgroups but showed weaker associations with the visuospatial and none subgroups.

Figure 6. Associations between identified AD risk factors and various AD subgroups, CSF biomarkers, and neuropathologic features. Asterisks denote significant associations based on an FDR cutoff of 0.05. P-values are truncated at 1e-5 for visualization purposes.



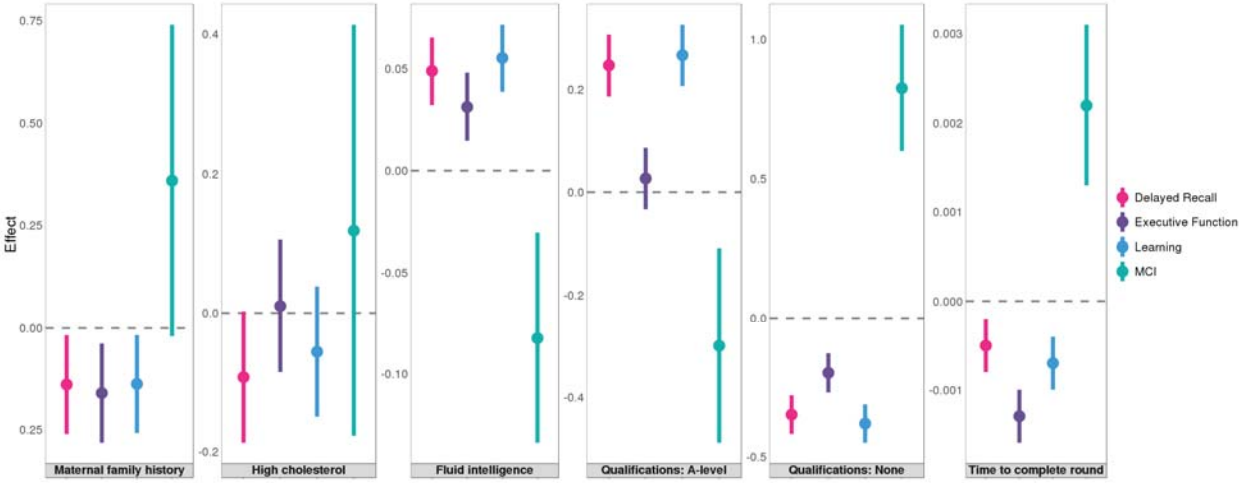
Next, we extended our analysis to three biomarkers of AD in cerebrospinal fluid (CSF): amyloid beta (Aβ_{42}), tau, and phosphorylated tau (ptau$_{181}$) [23]. Somewhat surprisingly, AD risk factors did not show strong associations with Aβ_{42} and tau (**Supplementary Table 13**). Maternal family history of AD and dementia was associated with ptau$_{181}$ (p=4.2e-4), but associations were absent for Aβ_{42} and tau. It has been recently suggested that CSF biomarkers have a sex-specific genetic architecture [24]. However, no association passed an FDR cutoff of 0.05 in our sex-stratified analyses (**Supplementary Table 14**).

Further, we applied BADGERS to a variety of neuropathologic features of AD and related dementias (**Methods**), including neuritic plaques (NPs), neurofibrillary tangles (NFTs), cerebral amyloid angiopathy (CAA), lewy body disease (LBD), hippocampal sclerosis (HS), and vascular brain injury (VBI) [25]. Family history of AD/dementia (p=3.8e-8, maternal; p=1.4e-5, paternal) and high cholesterol (p=2.1e-5) were strongly associated with NFT Braak stages (**Supplementary Table 15**). NP also showed very similar association patterns with these traits (p=2.7e-19, maternal family history; p=2.6e-7, paternal family history; p=0.001, high cholesterol). The other neuropathologic features did not show strong associations. Of note, despite not being statistically significant, family history of AD/dementia was negatively associated with VBI, and multiple intelligence traits were positively associated with LBD, showing distinct patterns with other pathologies (**Supplementary Figure 11**). We also note that various versions of the same pathologies all showed consistent associations in our analyses (**Supplementary Figure 11**). The complete association results for all the endophenotypes and all the traits are summarized in **Supplementary Table 16**. We further identified the influence of the *APOE* region in these results. The association results for all the endophenotypes with *APOE* Region being removed are summarized in **Supplementary Table 17**.

Associations with cognitive traits in a pre-clinical cohort

Finally, we studied the associations between AD risk factors and pre-clinical cognitive phenotypes using 1,198 samples from the Wisconsin Registry for Alzheimer's Prevention (WRAP), a longitudinal study of initially dementia-free middle-aged adults [26]. Assessed phenotypes include mild cognitive impairment (MCI) status and three cognitive composite scores for executive function, delayed recall, and learning (**Methods**). A total of 12 significant associations reached an FDR cutoff of 0.05 (**Supplementary Table 18**). Somewhat surprisingly, parental history and high cholesterol, the risk factors that showed the strongest associations with various AD endophenotypes, were not associated with MCI or cognitive composite scores in WRAP. Instead, education and intelligence-related traits strongly predicted pre-clinical cognition (**Figure 7**). A-levels education and no education both showed highly significant associations with delayed recall (p=4.0e-5 and 7.7e-7) and learning (p=7.6e-6 and 5.0e-8). No education was also associated with higher risk of MCI (p=2.5e-4). Additionally, fluid intelligence score was positively associated with the learning composite score (p=7.5e-4), and time to complete round in cognition test was negatively associated with the executive function (p=1.1e-5).

Figure 7. Associations between six traits and pre-clinical cognitive phenotypes in WRAP. Error bars denote the standard error of effect estimates.



Discussion

In this work, we introduced BADGERS, a new method to perform association scans at the biobank scale using genetic risk scores and GWAS association statistics. Through simulations, we demonstrated that our method provides consistent effect estimates and similar statistical power compared to regression analysis based on individual-level data. Additionally, we applied BADGERS to two large and independent GWAS datasets for late-onset AD. In our analyses, we used GWAS summary statistics from the UK biobank, one of the largest genetic cohort in the world, to generate PRS for complex traits. We estimated heritability for all available traits in the UK biobank and only included traits with nominally significant heritability ($p<0.05$) in our analyses. The GWAS summary statistics for Alzheimer's disease were also obtained from the largest available study – International Genomics of Alzheimer's Project (IGAP) and we further sought replication using a large, independent dataset from the Alzheimer's Disease Genetics Consortium (ADGC). Overall, we are confident that these quality control procedures largely controlled the false findings in our study. Among 1,738 heritable traits in the UK biobank, we identified 48 traits showing statistically significant associations with AD. These traits covered a variety of categories, including family history, cholesterol, intelligence, education, occupation, and lifestyle. Although many of the identified traits are genetically correlated, multivariate conditional analysis confirmed multiple strong and independent associations for AD. Family history showing strong associations with AD is not a surprise, and many other associations are supported by the literature as well. The protective effect of higher educational and occupational attainment on the risk and onset of dementia is well studied [27, 28]. Cholesterol buildup is also known to associate with β-amyloid plaques in the brain and higher AD risk [29-31].

More interestingly, these identified traits had distinct association patterns with various AD subgroups, biomarkers, pathologies, and pre-clinical cognitive traits. Five cognitively-defined AD subgroups were consistently associated with maternal family history, but only the group without substantial relative impairment in any domain (i.e. EPAD_none) was associated with intelligence and education. In addition, family history and high cholesterol were strongly associated with classic AD neuropathologies, including NP and NFT, while intelligence and educational attainment predicted pre-clinical cognitive scores and MCI. These results suggest that various AD risk factors may affect the disease course at different time points and via distinct biological processes, and genetically predicted risk factors for clinical AD include at least two separate components. While some risk factors (e.g. high cholesterol and *APOE*) may directly contribute to the accumulation of pathologies, other factors (e.g. intelligence and education) may buffer the adverse effect of brain pathology on cognition [28]. One possible scenario is that family history and high cholesterol are like the fundamental cause of AD while education level and intelligence are like the parameter of such factors. While if one didn't have such a factor in the first stage, they are protected from getting AD, if someone with such factor and also has high score in education attainment or intelligence, they can also get rid of the possibility of getting AD. We also investigated the influence of *APOE* on the identified associations. Effects of family history and high cholesterol were substantially reduced after *APOE* removal. In contrast, associations with cognition and education were virtually unchanged. These results suggest that various AD risk factors may affect the disease course at different time points and via distinct biological processes. While some risk factors (e.g. high cholesterol and *APOE*) may directly contribute to the accumulation of pathologies, other factors (e.g. intelligence and education) reduce the adverse effect of brain pathology on cognition [28].

Further, we note that the association results in BADGERS need to be interpreted with caution. Although PRS-based association analysis is sometimes treated as causal inference in the literature [32], we do not see BADGERS as a tool to identify causal factors. Key assumptions in causal inference are in many cases, violated when analyzing complex, highly polygenic traits, which may lead to complications when interpreting results. In our analysis, BADGERS showed superior statistical power than MR-IVW – among 1738 heritable traits, 48 reached Bonferroni significance in BADGERS, 9 and 18 traits reached Bonferroni significance in MR-IVW and GSMR, respectively. Among the 48 traits identified by BADGERS, 23 reached nominal statistical significance in MR-IVW and 11 were nominally significant in GSMR. BADGERS is a statistically powerful and computationally efficient method for identifying associations between a disease of interest and genetically imputed complex traits. Due to the capability of utilizing PRS with a large number of SNPs to impute complex traits, BADGERS has substantially improved statistical power compared to MR methods. And because of this, it can serve as a hypothesis-free method to screen for candidate risk factors from biobank-scale datasets with an overwhelming number of traits. After a list of candidate risk factors is identified using BADGERS, MR methods can be applied to carefully demonstrate causality. We envision BADGERS as a tool to prioritize associations among a large number of candidate risk factors so that robust causal inference methods can be applied to carefully assess causal effects. In addition, BADGERS requires a reference panel to provide LD estimates as a summary statistics-based method. If the population in the reference panel does not match that of the GWAS, it may create bias in the analysis. Our simulation results suggest that 1000 Genomes European samples is sufficient for our analysis when the GWAS was also conducted on European samples. Our implemented BADGERS software is flexible on the choice of LD reference panel. It allows users to change the reference dataset when they see fit.

What's more, environmental factors may play a big role in the identified associations. There is little doubt that environment could influence many complex traits, including the ones highlighted by the reviewer. However, this does not necessarily mean that these traits cannot also have a genetic component (or be genetically heritable). we summarized the heritability estimates for the 48 traits identified in our BADGERS meta-analysis of two independent datasets for Alzheimer's disease (**Supplementary Table 19**), and all of them have nominally significant heritability estimates ($p<0.05$) based on our selection criteria. Nevertheless, we do acknowledge that the high heritability of these traits is influenced by correlations with other traits. For example, job involving heavy manual or physical work is genetically correlated with educational attainment (Figure 3), which indicates that the association between this trait and Alzheimer's disease may not be direct. Therefore, it is important to note that association results from BADGERS analysis need to be interpreted with caution.

Limited sample size in AD endophenotypes is another limitation in our study. We have used data from the largest available GWAS for CSF biomarkers and neuropathologies. Still, small sample size made it challenging to assess the effects of traits that were weakly associated with AD. When an independent validation dataset is available, it would be of interest to assess the prediction accuracy of PRS on each trait. However, external validation datasets rarely exist in real applications. In that case, the users may choose to use heritability estimates to filter traits with a substantial genetic component. Further, in the BADGERS framework, PRS are independent variables in the regression analysis. If the PRS has limited predictive power, such noise is similar to measurement errors in standard regression analysis. This may decrease the statistical power in association tests but does not inflate the type-I error rate. Finally, emerging evidence has highlighted sex-specific genetic architecture of AD [24, 33]. In our analysis, maternal family history of AD showed stronger associations with various phenotypes than paternal family history. However, we note that this may be explained by the sample size difference in the UK biobank (N_{case}=28,507 and 15,022 for samples with maternal and paternal family history, respectively). We also performed sex-stratified analyses for CSF biomarkers but identified limited associations, possibly due to the small sample size. Overall, sex-specific effects of risk factors remain to be investigated in the future using larger datasets. In total, BADGERS requires the training data for genetic prediction models and the downstream disease GWAS to be independent but of similar genetic ancestry. Development of methods that are more robust to sample overlap and diverse genetic ancestry remains an open problem for future research.

In conclusion, BADGERS is a statistically powerful method to identify associated risk factors for complex diseases. Large-scale biobanks continue to provide rich data on various human traits that may be of interest in disease research. Our method uses GWAS to bridge large biobanks with studies on specific diseases, lessens the limitation of insufficient disease cases in biobanks and lack of risk factor measurements in disease studies, and provides a statistically-justified approach to identifying risk factors for disease. We have demonstrated the effectiveness of BADGERS through extensive simulations, a two-stage BWAS for late-onset AD, and various follow-up analyses on identified risk factors. Our results provided new insights into the genetic basis of AD, and revealed distinct mechanisms for the involvement of risk factors in AD etiologies. The ever-growing sample size in GWAS and biobanks, in conjunction with increasingly accessible summary association statistics, makes BADGERS a powerful and valuable tool in human genetics research.

Methods

BADGERS framework

The goal of this method is to study the association between Y, a measured trait in the study, and $\hat{T}$, a trait imputed from genetic data via a linear prediction model:

$$\hat{T} = XW$$

Here, $X_{N\times M}$ is the genotype matrix for N individuals in a study of trait Y. $W_{M\times 1}$ is the pre-calculated weight values on SNPs in the imputation model. M denotes the number of SNPs. We use Y, a $N \times 1$ vector, to denote the trait values measured on the same group of individuals. We test the association between Y and $\hat{T}$ via a linear model

$$Y = \alpha + \hat{T}\gamma + \delta$$

where α is the intercept, δ is the term for random noise, and regression coefficient γ is the parameter of interest. The ordinary least squares (OLS) estimator for γ can be denoted as

$$\hat{\gamma} = \frac{cov(\hat{T}, Y)}{var(\hat{T})} = \frac{cov(XW, Y)}{var(\hat{T})} = \frac{1}{var(\hat{T})} W^T \begin{pmatrix} cov(X_1, Y) \\ \vdots \\ cov(X_M, Y) \end{pmatrix}$$

Here, X_j is the jth column of X. Additionally, we derive the formula for the standard error of $\hat{\gamma}$:

$$se(\hat{\gamma}) = \sqrt{\frac{var(\delta)}{N \times var(\hat{T})}} \approx \sqrt{\frac{var(Y)}{N \times var(\hat{T})}}$$

The approximation in this formula is based on the assumption that trait Y has complex etiology and imputed trait $\hat{T}$ only explains a small proportion of its phenotypic variance. When an accurate estimate of $var(\delta)$ is difficult to obtain, this approximation approach provides conservative results and controls type-I error in the analysis.

In practice, individual-level genotype (i.e. X) and phenotype data (i.e. Y) may not be accessible due to policy and privacy concerns. Therefore, it is of practical interest to perform the aforementioned association analysis using summary association statistics. Standard genetic association analysis tests the association between trait Y and each SNP via the following linear model:

$$Y = \mu_j + X_j\beta_j + \varepsilon_j \ (1 \le j \le M)$$

The OLS estimator for β_j and its standard error have the following forms.

$$\hat{\beta}_j = \frac{cov(X_j, Y)}{var(X_j)}$$

$$se(\hat{\beta}_j) = \sqrt{\frac{var(\varepsilon_j)}{N \times var(X_j)}} \approx \sqrt{\frac{var(Y)}{N \times var(X_j)}}$$

Again, the approximation is based on the empirical observation in complex trait genetics – each SNP explains little variability of Y [34].

Next, we derive the test statistic (i.e. z-score) for γ:

$$\begin{aligned} Z &= \frac{\hat{\gamma}}{se(\hat{\gamma})} \\ &\approx \sqrt{\frac{N}{var(Y) \times var(\hat{T})}} W^T \begin{pmatrix} cov(X_1, Y) \\ \vdots \\ cov(X_M, Y) \end{pmatrix} \\ &\approx \frac{1}{\sqrt{var(\hat{T})}} W^T \begin{pmatrix} \frac{\sqrt{var(X_1)}\hat{\beta}_1}{se(\hat{\beta}_1)} \\ \vdots \\ \frac{\sqrt{var(X_M)}\hat{\beta}_M}{se(\hat{\beta}_M)} \end{pmatrix} \\ &= W^T \Gamma \tilde{Z} \end{aligned}$$

where Γ is a diagonal matrix with the jth diagonal element being

$$\Gamma_{jj} = \sqrt{\frac{var(X_j)}{var(\hat{T})}}$$

and $\tilde{Z}$ is the vector of SNP-level z-scores obtained from the GWAS of trait Y, i.e.

$$\tilde{Z}_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

Without access of individual-level genotype data, $var(X_j)$ and $var(\hat{T})$ need to be estimated using an external panel with a similar ancestry background. We use $\tilde{X}$ to denote the genotype matrix from an external cohort, then $var(X_j)$ can be approximated using the sample variance of $\tilde{X}_j$. Variance of $\hat{T}$ can be approximated as follows

$$var(\hat{T}) \approx W^T \widetilde{D} W$$

where $\widetilde{D}$ is the variance-covariance matrix of all SNPs estimated using $\tilde{X}$. However, when the number of SNPs is large in the imputation model for trait T, calculation of $\widetilde{D}$ is computationally intractable. Instead, we use an equivalent but computationally more efficient approach. We first impute trait T in the external panel using the same imputation model

$$\tilde{T} = \tilde{X}W$$

Then, $var(\hat{T})$ can be approximated by sample variance $var(\tilde{T})$.

Thus, we can test association between Y and $\hat{T}$ without having access to individual-level genotype and phenotype data from the GWAS. The required input variables for BADGERS include a linear imputation model for trait T, SNP-level summary statistics from a GWAS of trait Y, and an external panel of genotype data. With these, association test can be performed.

Multivariate analysis in BADGERS

To adjust for potential confounding effects, it may be of interest to include multiple imputed traits in the same BADGERS model. We still use Y to denote the measured trait of interest. The goal is to perform a multiple regression analysis using K imputed traits (i.e. $\hat{T}_1$, ..., $\hat{T}_K$) as predictor variables:

$$Y = \hat{T}^*\gamma^* + \delta^*$$

Here, we use $\hat{T}^* = (\hat{T}_1, \dots, \hat{T}_K)$ to denote a $N \times K$ matrix for K imputed traits. Regression coefficients $\gamma^* = (\gamma_1, \dots, \gamma_K)^T$ are the parameters of interest. To simplify algebra, we also assume trait Y and all SNPs in the genotype matrix X are centered so there is no intercept term in the model, but the conclusions apply to the general setting. Similar to univariate analysis, traits $\hat{T}_1, \dots, \hat{T}_K$ are imputed from genetic data via linear prediction models:

$$\hat{T}^* = XW^*$$

where $W^*_{M\times K}$ are imputation weights assigned to SNPs. The i$^{\text{th}}$ column of W denotes the imputation model for trait T_i. Then, the OLS estimator $\hat{\gamma}^*$ and its variance-covariance matrix can be denoted as follows:

$$\hat{\gamma}^* = ((\hat{T}^*)^T\hat{T}^*)^{-1}(\hat{T}^*)^TY$$

$$cov(\hat{\gamma}^*) \approx var(Y)((\hat{T}^*)^T\hat{T}^*)^{-1}$$

The approximation is based on the assumption that imputed traits $\hat{T}_1, \dots, \hat{T}_K$ collectively explain little variance in Y, which is reasonable in complex trait genetics if K is not too large. We further denote:

$$U := N\left((\hat{T}^*)^T\hat{T}^*\right)^{-1} = \begin{pmatrix} var(\hat{T}_1) & \cdots & cov(\hat{T}_1, \hat{T}_K) \\ \vdots & \ddots & \vdots \\ cov(\hat{T}_K, \hat{T}_1) & \cdots & var(\hat{T}_K) \end{pmatrix}^{-1}$$

All elements in matrix U can be approximated using a reference panel $\tilde{X}$ [35]:

$$cov(\hat{T}_i, \hat{T}_j) \approx cov(\tilde{T}_i, \tilde{T}_j)$$

Therefore, the z-score for γ_k $(1 \leq k \leq K)$ is

$$\begin{aligned} Z_k &= \frac{\hat{\gamma}_k}{se(\hat{\gamma}_k)} \\ &= \frac{I_k^T U (W^*)^T X^T Y}{\sqrt{N U_{kk} var(Y)}} \\ &= \frac{1}{\sqrt{U_{kk}}} I_k^T U (W^*)^T \Theta \tilde{Z} \end{aligned}$$

where I_k is the $K \times 1$ vector with the k^th^ element being 1 and all other elements equal to 0, Θ is a $M \times M$ diagonal matrix with the i^th^ diagonal element being $\sqrt{var(X_i)}$, and similar to the notation in univariate analysis, $\tilde{Z}$ is the vector of SNP-level z-scores from the GWAS of trait Y. Given imputation models for K traits (i.e. W^*), GWAS summary statistics for trait Y (i.e. $\tilde{Z}$), and an external genetic dataset to estimate U and Θ, multivariate association analysis can be performed without genotype and phenotype data from the GWAS.

Genetic prediction

Any linear prediction model can be used in the BADGERS framework. With access to individual-level genotype and phenotype data, the users can train their preferred statistical learning models, e.g. penalized regression or linear mixed model. When only GWAS summary statistics are available for risk factors (i.e. T), PRS can be used for imputation. We used PRS to impute complex traits in all analyses throughout the paper. Of note, more advanced PRS methods that explicitly model LD [36] and functional annotations [37] to improve prediction accuracy have been developed. However, additional independent datasets may be needed if there are tuning parameters in PRS. In general, higher imputation accuracy will improve statistical power in association testing [12]. The BADGERS software allows users to choose their preferred imputation model.

Simulation settings

We simulated quantitative traits using genotype data of 62,313 individuals from the GERA cohort (dbGap accession: phs000674). Summary association statistics were generated using PLINK [38]. We ran BADGERS on summary statistics based on the simulated traits and PRS of 1,738 traits in the UK biobank. To compare BADGERS with the traditional approach that uses individual-level data as input, we also directly regressed simulated traits on the PRS of UK biobank traits to estimate association effects.

Setting 1. We simulated quantitative trait values as i.i.d. samples from normal distribution with mean 0 and variance 1. In this setting, simulated trait values were independent from genotype data.

Setting 2. We simulated quantitative trait values based on an additive random effect model commonly used in heritability estimation [39]. We fixed heritability to be 0.1. In this setting, the simulated trait is associated with SNPs, but is not directly related to PRS of UK biobank traits.

Setting 3. We selected 100 traits from 1,738 UK-Biobank traits to calculate PRS on GERA data. For each of these 100 PRS, we simulated a quantitative trait by summing up the effect of PRS, a polygenic genetic background, and a noise term.

$$Y = X\beta + \rho P + \varepsilon$$

Here, X denotes the genotype of samples; β is the effect size of each variant; P is the PRS of one of the selected traits; ρ is the effect size of PRS; and ε is the error term following a standard normal distribution. The polygenic background and random noise (i.e. $X\beta + \varepsilon$) were simulated using the same model described in setting 2. This term and the PRS were normalized separately. The standardized effect size (i.e. ρ) was set as 0.02, 0.015, 0.01, 0.008, and 0.005 in our simulations. In this setting, simulated traits are directly associated with SNPs and PRS. For each value of ρ, statistical power was calculated as the proportion of significant results (p<0.05) out of 100 traits.

Setting 4. We simulated 100 quantitative traits $T_1, \ldots, T_{100}$ based on an additive random effect model commonly used with heritability fixed as 0.1. And the response traits $Y_1, \ldots, Y_{100}$ were simulated by adding a noise term to T.

$$Y_i = \gamma_i T_i + \varepsilon_i$$

Where $\gamma_i \sim N(0,2)$, and $\varepsilon_i \sim N\left(0, Var(T_i)\right)$. The data set was split into two subsets, one with 31,162 (subset 1) and another with 31,163 samples (subset 2). Marginal summary statistics correspond to T_i's and Y_i's were derived using subset 1 and subset 2, respectively. We applied LDpred to jointly estimate all SNPs' effects using marginal summary statistics from subset 1.

Then, we ran BADGERS to identify associations between 100 pairs of Y_i and T_i using two methods to impute T_i's (i.e. marginal PRS and LDpred).

GWAS datasets

Summary statistics for 4,357 UK biobank traits were generated by Dr. Benjamin Neale's group and were downloaded from (http://www.nealelab.is/uk-biobank). AD summary statistics from the IGAP stage-I analysis were downloaded from the IGAP website (http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php). ADGC phase 2 summary statistics were generated by first analyzing individual datasets using logistic regression adjusting for age, sex and the first three principal components in the program SNPTest v2 [40]. Meta-analysis of the individual dataset results was then performed using the inverse-variance weighted approach [41].

GWAS summary statistics for neuropathologic features of AD and related dementias were obtained from the ADGC. Details on these data have been previously reported [25]. We analyzed a total of 13 neuropathologic features, including four NP traits, two traits for NFT Braak stages, three traits for LBD, CAA, HS, and two VBI traits. Among different versions of the same pathology, we picked one dataset for each pathologic feature to show in our primary analyses. Six AD subgroups were defined in the recent EPAD paper [21] on the basis of relative performance in memory, executive functioning, visuospatial functioning, and language at the time of Alzheimer's diagnosis. Four subgroups include AD samples with an isolated substantial relative impairment in one of four domains; the "none" subgroup includes samples without substantial relative impairment; the "mix" subgroup includes samples with relative impairment in multiple domains. Each domain was compared with healthy controls in case-control association analyses. We did not include the executive functioning subgroup in our analysis due to its small sample size in cases. Detailed information about the design of CSF biomarker GWAS and the recent sex-stratified analysis has been described previously [23, 24]. Details on the association statistics for AD subgroups, CSF biomarkers, and neuropathological features are summarized in **Supplementary Table 17**.

Analysis of GWAS summary statistics

We applied LD score regression implemented in the LDSC software [42] to estimate the heritability of each trait. Among 4,357 traits, we selected 1,738 with nominally significant heritability (p<0.05) to include in our analyses. We removed SNPs with association p-values greater than 0.01 from each of the 1,738 summary statistics files, clumped the remaining SNPs using a LD cutoff of 0.1 and a radius of 1 Mb in PLINK [38], and built PRS for each trait using the effect size estimates of remaining SNPs.

Throughout the paper, we used samples of European ancestry in the 1000 Genomes Project as a reference panel to estimate LD [43]. In univariate analyses, we tested marginal associations between each PRS and AD using the IGAP stage-I dataset and replicated the findings using the ADGC summary statistics. Association results in two stages were combined using an inverse variance-weighted meta-analysis [41]. A stringent Bonferroni-corrected significance threshold was used to identify AD-associated risk factors. For associations between identified risk factors and AD endophenotypes, we used an FDR cutoff of 0.05 to claim statistical significance. We applied hierarchical clustering to the covariance of 48 traits we identified from marginal association analysis, then divided the result into 15 clusters and selected one most significant trait from each cluster and used them to perform multivariate conditional analysis. We analyzed IGAP and ADGC datasets separately, and combined the results using meta-analysis.

We used MR-IVW approach [44] implemented in the Mendelian Randomization R package [45] to study the causal effects of 48 risk factors identified by BADGERS. For each trait, we selected instrumental SNP variables as the top 30 most significant SNPs after clumping all SNPs using a LD cutoff of 0.1.

Analysis of WRAP data

WRAP is a longitudinal study of initially dementia-free middle-aged adults that allows for the enrollment of siblings and is enriched for a parental history of AD. Details of the study design and methods used have been previously described [26, 46]. After quality control, a total of 1,198 participants whose genetic ancestry was primarily of European descent were included in our analysis. On average, participants were 53.7 years of age (SD=6.6) at baseline and had a bachelor's degree, and 69.8% (n=836) were female. Participants had two to six longitudinal study visits, with an average of 4.3 visits, leading to a total of 5,184 observations available for analyses.

DNA samples were genotyped using the Illumina Multi-Ethnic Genotyping Array at the University of Wisconsin Biotechnology Center. Thirty-six blinded duplicate samples were used to calculate a concordance rate of 99.99%, and discordant genotypes were set to missing. Imputation was performed with the Michigan Imputation Server v1.0.3 [47], using the Haplotype Reference Consortium (HRC) v. r1.1 2016 [48] as the reference panel and Eagle2 v2.3 [49] for phasing. Variants with a quality score $R^2<0.80$, MAF<0.001, or that were out of HWE were excluded, leading to 10,499,994 imputed and genotyped variants for analyses. Data cleaning and file preparation were completed using PLINK v1.9 [50] and VCFtools v0.1.14 [51]. Coordinates are based on the hg19 genome build. Due to the sibling relationships present in the WRAP cohort, genetic ancestry was assessed and confirmed using Principal Components Analysis in Related Samples (PC-AiR), a method that makes robust inferences about population structure in the presence of relatedness [52].

Composite scores were calculated for executive function, delayed recall, and learning based on a previous analysis [53]. Each composite score was calculated from three neuropsychological tests, which were each converted to z-scores using baseline means and standard deviations. These z-scores were then averaged to derive executive function and delayed recall composite scores at each visit for each individual. Cognitive impairment status was determined based on a consensus review by a panel of dementia experts. Resulting cognitive statuses included cognitively normal, early MCI, clinical MCI, impairment that was not MCI, or dementia, as previously defined [54]. Participants were considered cognitively impaired if their worst consensus conference diagnosis was early MCI, clinical MCI, or dementia (n=387). Participants were considered cognitively stable if their consensus conference diagnosis was cognitively normal across all visits (n=803).

The 48 PRSs were developed within the WRAP cohort using PLINK v1.9 [50] and tested for associations with the three composite scores (i.e. executive function, delayed recall, and learning) and cognitive impairment statuses. MCI status was tested using logistic regression models in R, while all other associations, which utilized multiple study visits, were tested using linear mixed regression models implemented in the lme4 package in R [55]. All models included fixed effects for age and sex, and cognitive composite scores additionally included a fixed effect for practice effect (using visit number). Mixed models included random intercepts for withinsubject correlations due to repeated measures and within-family correlations due to the enrollment of siblings.

Software availability

The BADGERS software is freely available at https://github.com/qlu-lab/BADGERS

Acknowledgments

This project was supported by the Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427. Support for this research was also provided by the University of Wisconsin-Madison Office of the Chancellor and the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. BFD was supported by an NLM training grant to the Computation and Informatics in Biology and Medicine Training Program [NLM 5T15LM007359]. This research was also supported by the NIH [grants R01AG054047, R01AG27161, UL1TR000427, and P2C HD047873], Helen Bader Foundation, Northwestern Mutual Foundation, Extendicare Foundation, and State of Wisconsin. The authors thank the University of Wisconsin Madison Biotechnology Center Gene Expression Center for providing Illumina Infinium genotyping services. We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the subjects and their families. The i-Select chips were funded by the French National Foundation on Alzheimer's disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Université de Lille 2, and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant n° 503480), Alzheimer's Research UK (Grant n° 503176), the Wellcome Trust (Grant n° 082604/2/07/Z), and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant n° 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01–AG–12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer's Association grant ADGC–10–196728. We thank contributors who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible; Data for this study were prepared, archived, and distributed by the National Institute on Aging Alzheimer's Disease Data Storage Site (NIAGADS) at the University of Pennsylvania (U24-AG041689-01). We are also grateful for ADGC and its investigators for providing GWAS summary statistics for various AD phenotypes. The full acknowledgement to ADGC is included in the supplementary material.

Competing financial interests

The authors declare no competing financial interests.

Author contribution

Q.L. conceived the study.

D.Y., B.H., and Q.L. developed the statistical framework.

D.Y., B.H., B.D., and Q.L. performed the statistical analyses.

D.Y. implemented the software.

S.M., B.W.K., Y.D., L.D., A.N., A.K., Y.Z., C.C., and T.J.H., assisted in preparation and curation of GWAS association statistics.

S.M. and P.K.C. assisted in data application and interpretation of association results.

B.F.D., S.C.J., and C.D.E. assisted in WRAP data preparation and interpretation.

Y.W. assisted in literature review and results interpretation.

Q.L. advised on statistical and genetic issues.

H.K. advised on causal inference.

D.Y. B.H., and Q.L. wrote the manuscript.

All authors contributed in manuscript editing and approved the manuscript.

References

1. Prince, M., et al., *The global prevalence of dementia: a systematic review and metaanalysis.* Alzheimer's & Dementia, 2013. **9**(1): p. 63-75. e2.
2. Reitz, C. and R. Mayeux, *Alzheimer disease: epidemiology, diagnostic criteria, risk factors and biomarkers.* Biochemical pharmacology, 2014. **88**(4): p. 640-651.
3. Østergaard, S.D., et al., *Associations between potentially modifiable risk factors and Alzheimer disease: a Mendelian randomization study.* PLoS medicine, 2015. **12**(6): p. e1001841.
4. Larsson, S.C., et al., *Modifiable pathways in Alzheimer's disease: Mendelian randomisation analysis.* bmj, 2017. **359**: p. j5375.
5. Norton, S., et al., *Potential for primary prevention of Alzheimer's disease: an analysis of population-based data.* The Lancet Neurology, 2014. **13**(8): p. 788-794.
6. Jack Jr, C.R., et al., *Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers.* The Lancet Neurology, 2013. **12**(2): p. 207-216.
7. Sleiman, P.M.A. and S.F.A. Grant, *Mendelian Randomization in the Era of Genomewide Association Studies.* Clinical Chemistry, 2010. **56**(5): p. 723.
8. Davey Smith, G. and G. Hemani, *Mendelian randomization: genetic anchors for causal inference in epidemiological studies.* Human Molecular Genetics, 2014. **23**(R1): p. R89-R98.
9. Zhu, Z., et al., *Causal associations between risk factors and common diseases inferred from GWAS summary data.* Nature Communications, 2018. **9**(1): p. 224.
10. Gamazon, E.R., et al., *A gene-based association method for mapping traits using reference transcriptome data.* Nature genetics, 2015. **47**(9): p. 1091-1098.
11. Gusev, A., et al., *Integrative approaches for large-scale transcriptome-wide association studies.* Nature genetics, 2016. **48**(3): p. 245-252.
12. Hu, Y., et al., *A statistical framework for cross-tissue transcriptome-wide association analysis.* bioRxiv, 2018: p. 286013.

13. Lambert, J.-C., et al., *Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.* Nature genetics, 2013. **45**(12): p. 1452-1458.

14. Harold, D., et al., *Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease.* Nature genetics, 2009. **41**(10): p. 1088-1093.

15. Hollingworth, P., et al., *Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease.* Nature genetics, 2011. **43**(5): p. 429-435.

16. Naj, A.C., et al., *Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease.* Nature genetics, 2011. **43**(5): p. 436-441.

17. Seshadri, S., et al., *Genome-wide analysis of genetic loci associated with Alzheimer disease.* Jama, 2010. **303**(18): p. 1832-1840.

18. Jun, G.R., et al., *Transethnic genome-wide scan identifies novel Alzheimer's disease loci.* Alzheimer's & Dementia, 2017.

19. Bycroft, C., et al., *Genome-wide genetic data on~ 500,000 UK Biobank participants.* bioRxiv, 2017: p. 166298.

20. Vilhjálmsson, Bjarni J., et al., *Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores.* The American Journal of Human Genetics, 2015. **97**(4): p. 576-592.

21. Mukherjee, S., et al., *Genetic data and cognitively-defined late-onset Alzheimer's disease subgroups.* bioRxiv, 2018.

22. Crane, P.K., et al., *Incidence of cognitively defined late-onset Alzheimer's dementia subgroups from a prospective cohort study.* Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 2017. **13**(12): p. 1307-1316.

23. Deming, Y., et al., *Genome-wide association study identifies four novel loci associated with Alzheimer's endophenotypes and disease modifiers.* Acta neuropathologica, 2017. **133**(5): p. 839-856.

24. Deming, Y., et al., *Sex-specific genetic predictors of Alzheimer's disease biomarkers.* Acta Neuropathologica, 2018.

25. Beecham, G.W., et al., *Genome-Wide Association Meta-analysis of Neuropathologic Features of Alzheimer's Disease and Related Dementias.* PLOS Genetics, 2014. **10**(9): p. e1004606.

26. Johnson, S.C., et al., *The Wisconsin Registry for Alzheimer's Prevention: A review of findings and current directions.* Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 2018. **10**: p. 130-142.

27. Valenzuela, M.J. and P. Sachdev, *Brain reserve and dementia: a systematic review.* Psychological medicine, 2006. **36**(4): p. 441-454.

28. Stern, Y., *Cognitive reserve in ageing and Alzheimer's disease.* The Lancet Neurology, 2012. **11**(11): p. 1006-1012.

29. Reed, B., et al., *Associations between serum cholesterol levels and cerebral amyloidosis.* JAMA neurology, 2014. **71**(2): p. 195-200.

30. Djelti, F., et al., *CYP46A1 inhibition, brain cholesterol accumulation and neurodegeneration pave the way for Alzheimer's disease.* Brain, 2015. **138**(8): p. 2383-2398.

31. Simons, M., et al., *Cholesterol and Alzheimer's disease Is there a link?* Neurology, 2001. **57**(6): p. 1089-1093.

32. Paternoster, L., K. Tilling, and G.D. Smith, *Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges.* PLoS genetics, 2017. **13**(10): p. e1006944.

33. Hohman, T.J., et al., *Sex-Specific Association of Apolipoprotein E With Cerebrospinal Fluid Levels of Tau.* JAMA neurology, 2018.

34. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-753.

35. Dudbridge, F., *Power and Predictive Accuracy of Polygenic Risk Scores.* PLOS Genetics, 2013. **9**(3): p. e1003348.

36. Vilhjálmsson, B.J., et al., *Modeling linkage disequilibrium increases accuracy of polygenic risk scores.* The American Journal of Human Genetics, 2015. **97**(4): p. 576-592.

37. Hu, Y., et al., *Leveraging functional annotations in genetic risk prediction for human complex diseases.* PLOS Computational Biology, 2017. **13**(6): p. e1005589.

38. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

39. Yang, J., et al., *Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index.* Nature genetics, 2015. **47**(10): p. 1114-1120.

40. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nature genetics, 2007. **39**(7): p. 906.

41. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics, 2010. **26**(17): p. 2190-1.

42. Bulik-Sullivan, B.K., et al., *LD Score regression distinguishes confounding from polygenicity in genome-wide association studies.* Nature Genetics, 2015. **47**: p. 291.

43. Abecasis, G.R., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

44. Burgess, S., A. Butterworth, and S.G. Thompson, *Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data.* Genetic Epidemiology, 2013. **37**(7): p. 658-665.

45. Yavorska, O.O. and S. Burgess, *MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data.* International Journal of Epidemiology, 2017. **46**(6): p. 1734-1739.

46. Sager, M.A., B. Hermann, and A. La Rue, *Middle-aged children of persons with Alzheimer's disease: APOE genotypes and cognitive function in the Wisconsin Registry for Alzheimer's Prevention.* J Geriatr Psychiatry Neurol, 2005. **18**(4): p. 245-9.

47. Das, S., et al., *Next-generation genotype imputation service and methods.* Nat Genet, 2016. **48**(10): p. 1284-1287.

48. McCarthy, S., et al., *A reference panel of 64,976 haplotypes for genotype imputation.* Nat Genet, 2016. **48**(10): p. 1279-83.

49. Loh, P.R., et al., *Reference-based phasing using the Haplotype Reference Consortium panel.* Nat Genet, 2016. **48**(11): p. 1443-1448.

50. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**: p. 7.

51. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.

52. Conomos, M.P., M.B. Miller, and T.A. Thornton, *Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness.* Genet Epidemiol, 2015. **39**(4): p. 276-93.

53. Clark, L.R., et al., *Beta-amyloid and cognitive decline in late middle age: Findings from the Wisconsin Registry for Alzheimer's Prevention study.* Alzheimers Dement, 2016. **12**(7): p. 805-14.

54. Koscik, R.L., et al., *Intraindividual Cognitive Variability in Middle Age Predicts Cognitive Impairment 8-10 Years Later: Results from the Wisconsin Registry for Alzheimer's Prevention.* J Int Neuropsychol Soc, 2016. **22**(10): p. 1016-1025.

55. Bates, D., et al., *Fitting Linear Mixed-Effects Models Using lme4.* Journal of Statistical Software, 2015. **67**(1): p. 1-48.